VideoGPT: Video Generation using VQ-VAE and Transformers

Year: 2021 | Citations: 622 | Authors: Wilson Yan, Yunzhi Zhang, P. Abbeel, A. Srinivas

Abstract

We present VideoGPT: a conceptually simple architecture for scaling likelihood based generative modeling to natural videos. VideoGPT uses VQ-VAE that learns downsampled discrete latent representations of a raw video by employing 3D convolutions and axial self-attention. A simple GPT-like architecture is then used to autoregressively model the discrete latents using spatio-temporal position encodings. Despite the simplicity in formulation and ease of training, our architecture is able to generate samples competitive with state-of-the-art GAN models for video generation on the BAIR Robot dataset, and generate high fidelity natural videos from UCF-101 and Tumbler GIF Dataset (TGIF). We hope our proposed architecture serves as a reproducible reference for a minimalistic implementation of transformer based video generation models. Samples and code are available at https://wilson1yan.github.io/videogpt/index.html